University of North Texas at Dallas

Strategic Analysis & Reporting

A Predictive Model for Student Retention Using Logistic Regression

1. **Abstract**

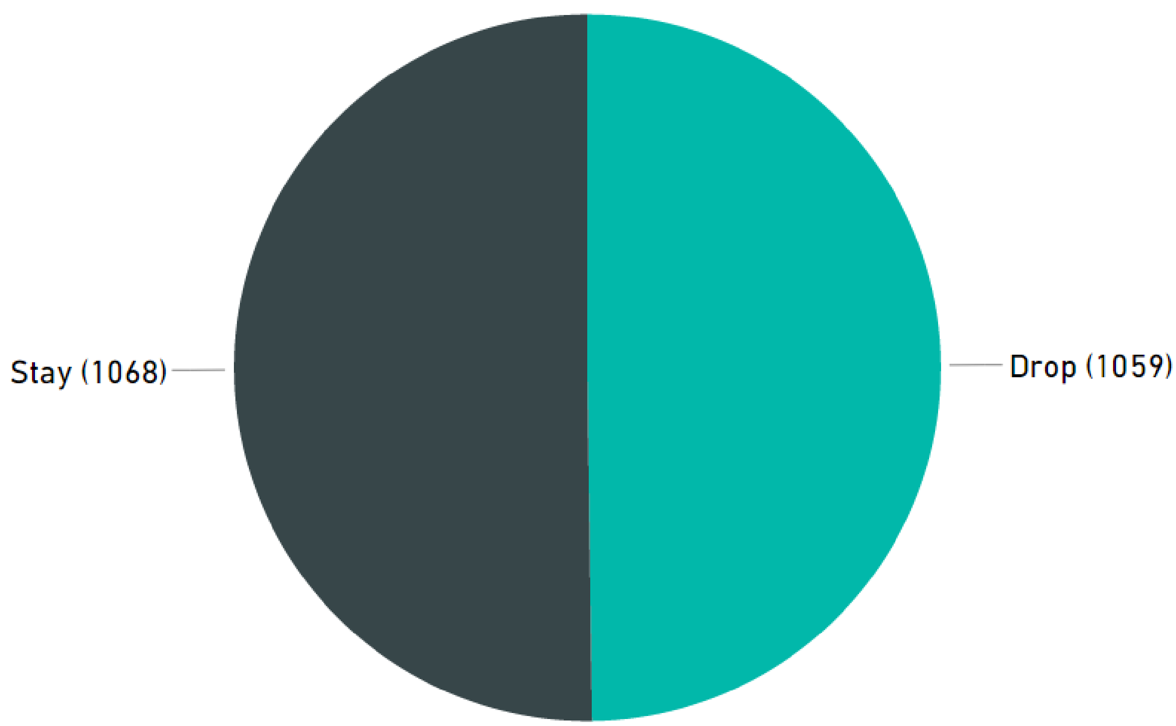
The percentage of students in a university or college who return to the institution after one year's study (called retention) is crucial for decision makers since it is one of the performance measures of higher education institutions. The decision makers would know how well the institution supports students who have academic, financial, and/or other challenges. It provides a window into different aspects of the institution. In addition, College-to-be students use retention to make college choice decisions. Hence, retention is an important measurement for decision makers to decide on recruitment policies.

With the purpose to know which variables influence the students' retention at UNT Dallas, we created a model using logistic regression to compare impacts of variables on the retention. In particular, we focused on how the selected variables influence the retention as well as the relationships between retention and these variables.

2. **Theoretical framework**

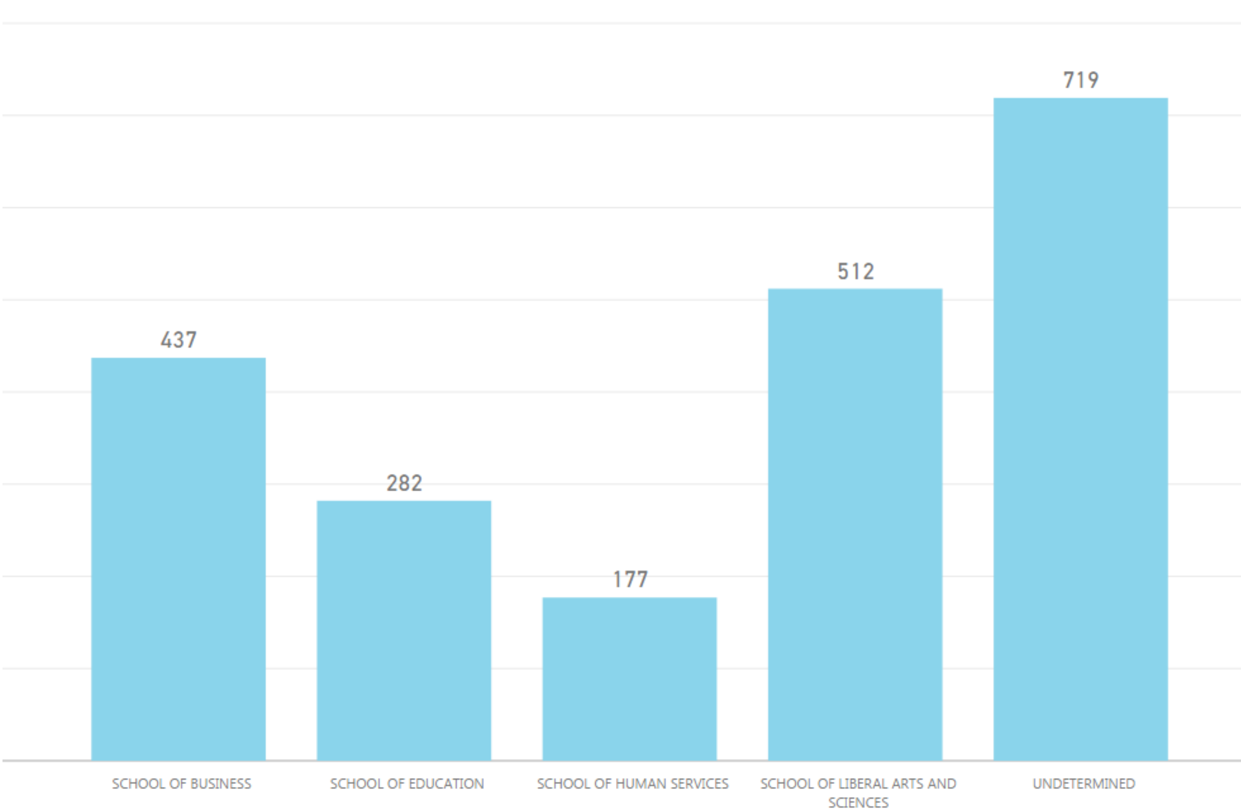
Data preparation

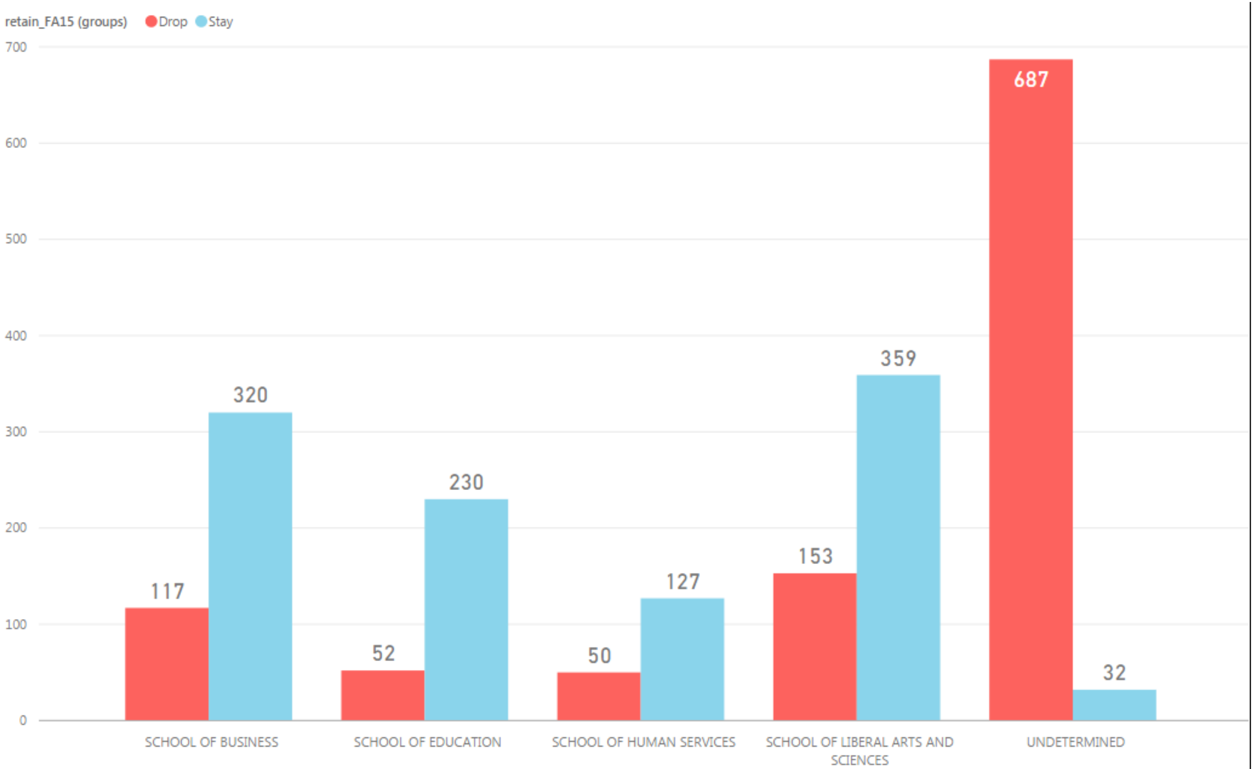
We selected the students, only undergraduate, who attended UNT Dallas in 2014 fall as our sample dataset. Then we eliminated the students who graduated between 2014 fall and 2015 fall and all senior students. Lastly, we compared this dataset with the students who attended UNT Dallas in 2015 fall. We marked a student as 1 if the student returned to UNT Dallas in 2015 fall and 0 for those who did not.



As the graph shows above, there were a total of 2127 undergraduate students (senior excluded). More students stayed than those who dropped, 1068 compared with 1059.

Chart 2 School or College

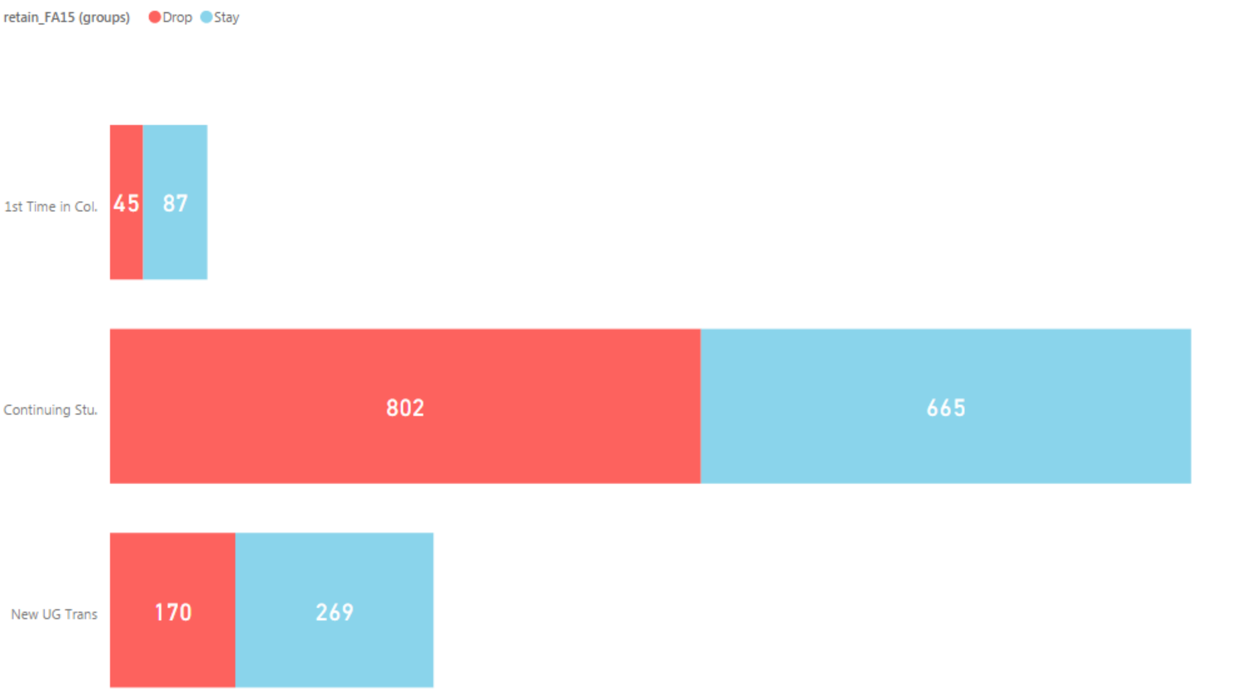




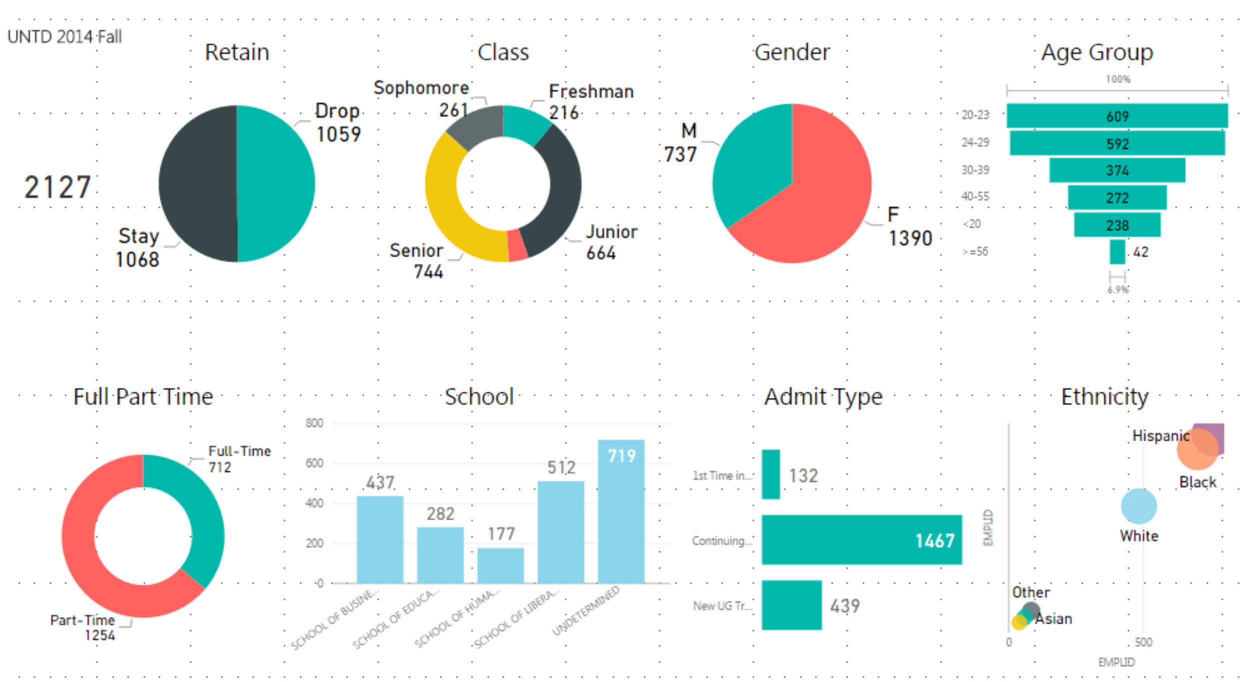
As chart two shows, the sample dataset covers all the colleges exclude the College of Law. The remarkable phenomenon that draws our attention is the retention rate of undetermined students, which is extremely low compared to other colleges. Further research will be done in order to identify the impacts of Undetermined. However, we removed the Undetermined from the model for outliner consideration.

Also, the distribution of the students among colleges in the sample dataset is very similar to the distribution that we have in population.

Chart 3 Admit Type

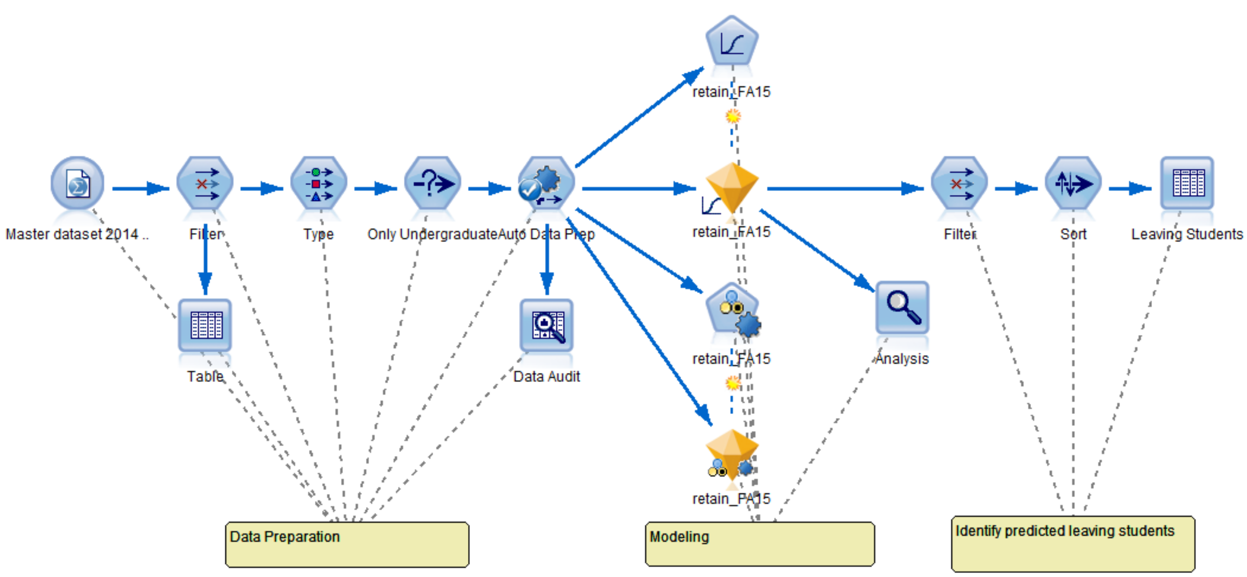


As the chart three shows, less student returned to UNT Dallas if they have been classified as continuing students.

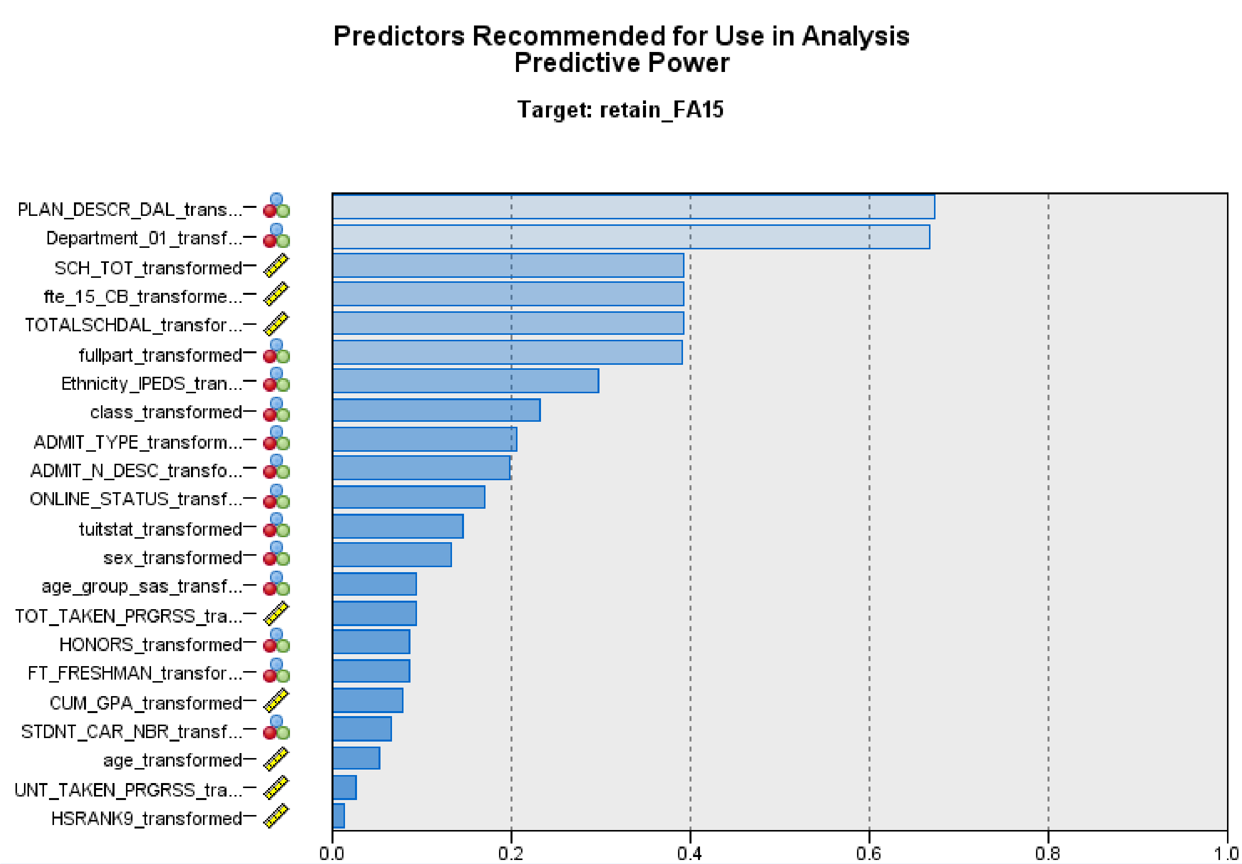


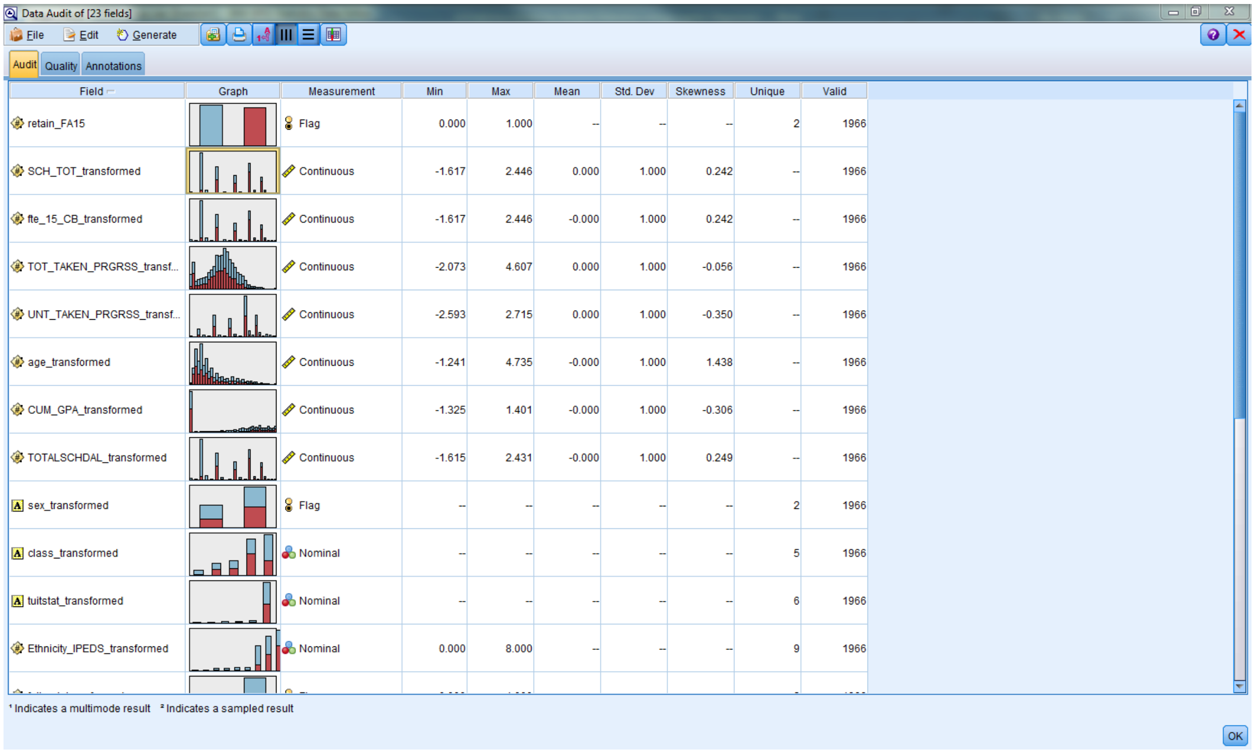
The chart above shows different dimensions of the students in our dataset.

Model Building



The software that we used in the model building is SPSS Modeler and SPSS Statistics. The algorithm used in this predictive model is <u>Logistic Regression</u>, sometimes called Binary Logistic. <u>CHAID</u>, and <u>NN</u>. In the modeling, the target column is <u>Retain</u>, a categorical variable. And we select 22 independent variables.

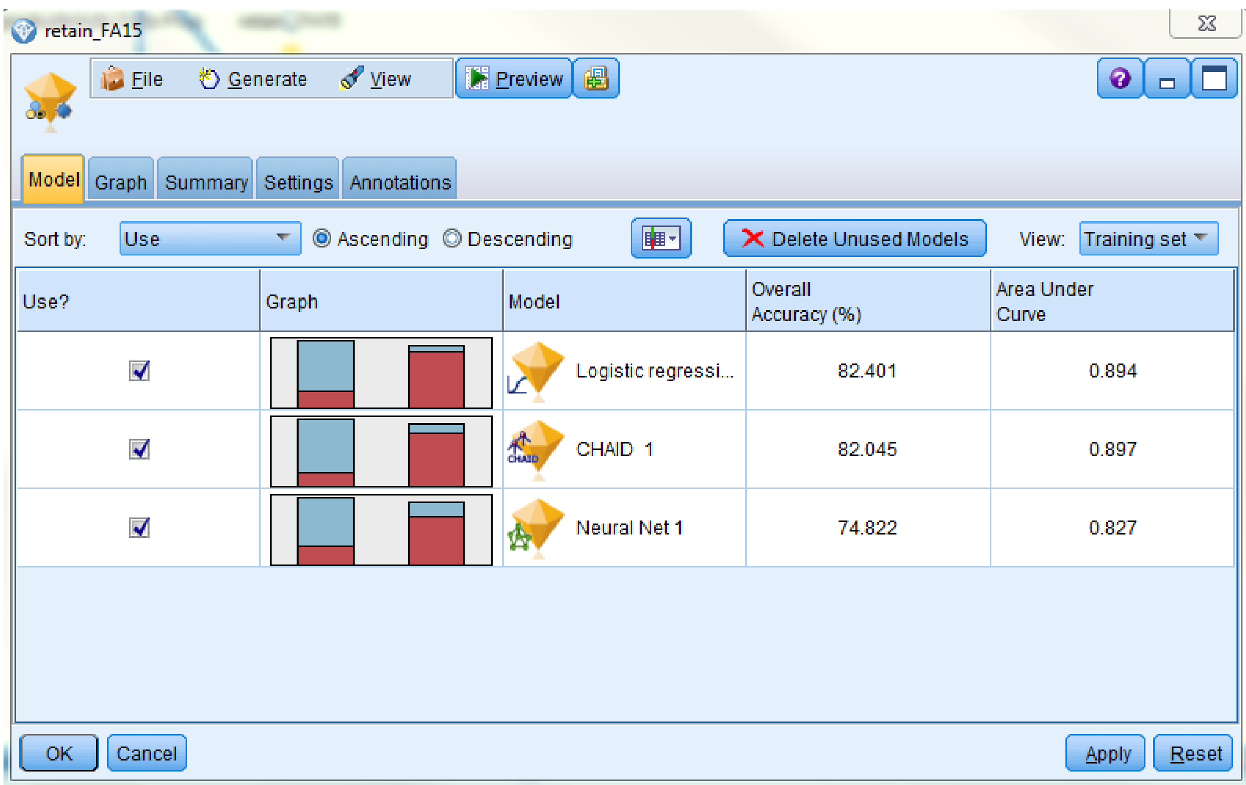




Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
retain_FA15		Flag	0.000	1.000	--	--	--	2	1966
SCH_TOT_transformed		Continuous	-1.617	2.446	0.000	1.000	0.242	--	1966
fte_15_CB_transformed		Continuous	-1.617	2.446	-0.000	1.000	0.242	--	1966
TOT_TAKEN_PRGRSS_transf...		Continuous	-2.073	4.607	0.000	1.000	-0.056	--	1966
UNT_TAKEN_PRGRSS_transf...		Continuous	-2.593	2.715	0.000	1.000	-0.350	--	1966
age_transformed		Continuous	-1.241	4.735	-0.000	1.000	1.438	--	1966
CUM_GPA_transformed		Continuous	-1.325	1.401	-0.000	1.000	-0.306	--	1966
TOTALSCHDAL_transformed		Continuous	-1.615	2.431	-0.000	1.000	0.249	--	1966
sex_transformed		Flag	--	--	--	--	--	2	1966
class_transformed		Nominal	--	--	--	--	--	5	1966
tuitstat_transformed		Nominal	--	--	--	--	--	6	1966
Ethnicity_IPEDS_transformed		Nominal	0.000	8.000	--	--	--	9	1966

[1] Indicates a multimode result [2] Indicates a sampled result

3. **Results and Conclusions**



Use?	Graph	Model	Overall Accuracy (%)	Area Under Curve
☑		Logistic regressi...	82.401	0.894
☑		CHAID 1	82.045	0.897
☑		Neural Net 1	74.822	0.827

As the charts shows above. Both three models generate decent accuracy rate, Logistic Regression is the highest, 82.4%, followed by CHAID, 82%, and Neural Net, 74.8%.

In the future usage of the model, we could predict whether a student will retain. We could also interpret the result of Logistic Regression and understand which variables is important for us. In the following chart, we selected the most significant variables.

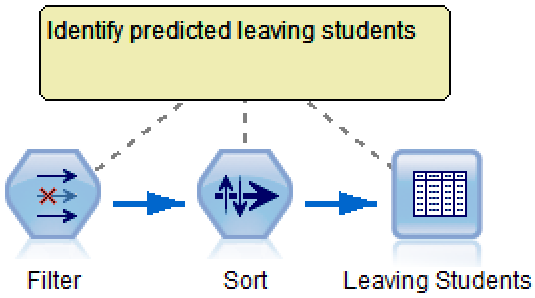
Variable	Category	B	P	Exp(B)
PLAN_DESCR_DAL	Non-Degree	-2	0.001	0.13
	Criminal Justice	1.1	0	3
	Interdisciplinary Studies	1.3	0	4
	Business	1	0	3
CLASS	Freshmen	-1.3	0.003	0.3
	Sophomore	-0.6	0.013	0.5
TOT_TAKEN_PRGRSS		-0.6	0	0.5
AGE		0.5	0.48	1.7
CUM_GPA		0.5	0.001	1.6
ONLINE_STATUS	Both	-1.3	0.46	0.3

As the results show, students who have been categorized as Non-Degree have less chance to retain. And students who have major of Criminal Justice, Interdisciplinary Studies, and Business have higher chance to retain. Freshmen and Sophomore students are less likely to retain. And the more SCH a student takes during that semester, the more likely he/she will drop. Age indicates the older the student is, the more likely he/she will retain. Cumulative GPA has a positive relationship with retention, the higher the cumulative GPA is, the more likely the student will retain. Another interesting finding is Online status, compare to students who only take online classes and those who only take on-campus classes, students who take both online and on-campus classes have lower chance to retain.

Hence, the model is appropriate to predict students' retention by using all the variables.

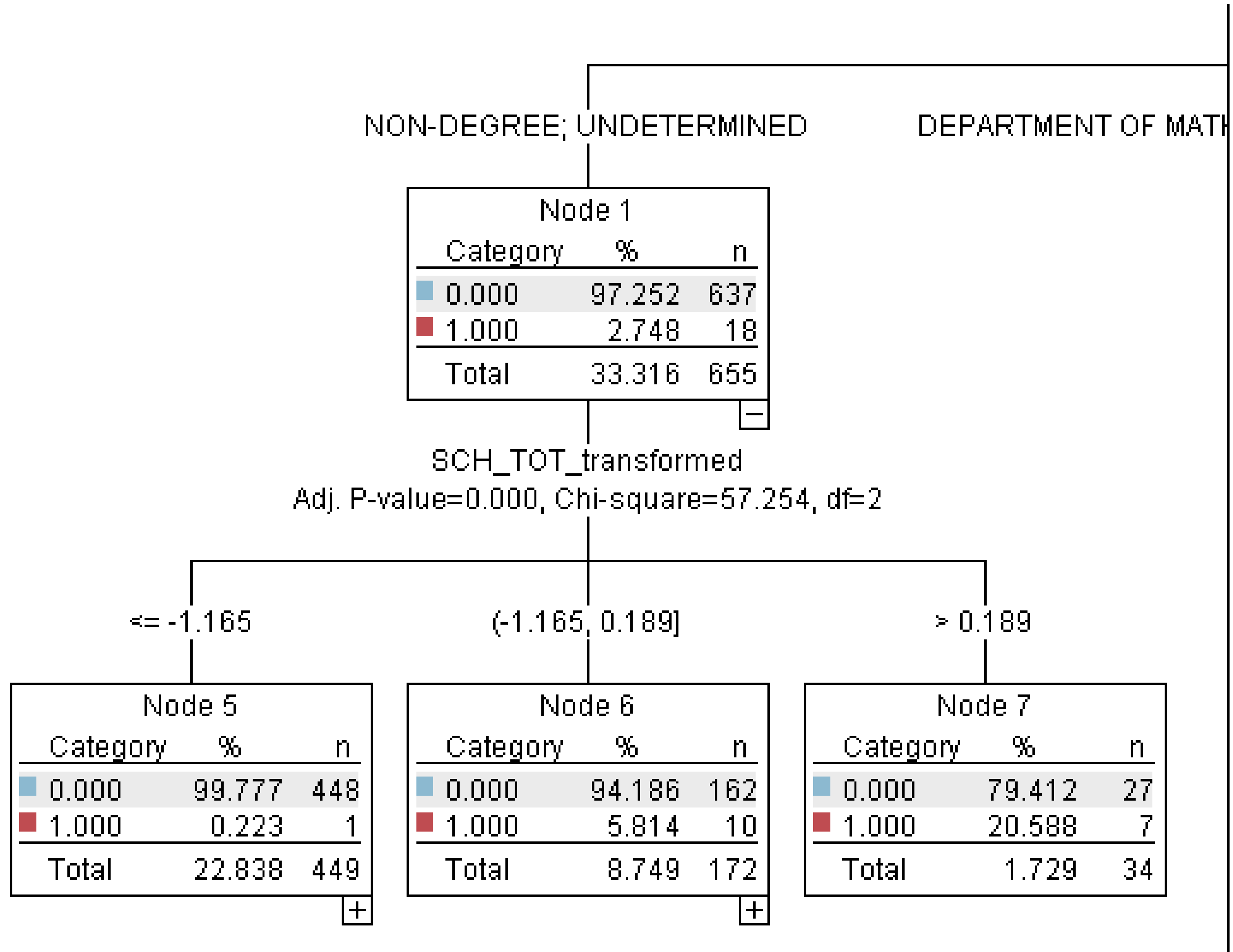
4. **Usage of the results and Future Research**

With a predictive model of 82.4% accuracy rate, we can use it in the first semester to predict who will be most likely to drop in the next semester. By the possibility index the system generates, we can rank the students from highest possibility of dropping to lowest. With the list, we can contact those students who are most likely to drop and offer some help or intervention to help them to come back next semester.



	EMPLID	$LP-0
1	10578844	1.000
2	10796709	1.000
3	10953307	1.000
4	10885715	1.000
5	10890892	1.000
6	10956116	1.000
7	10917726	1.000
8	10423011	1.000
9	10826935	1.000
10	11049815	1.000
11	11029257	1.000
12	10958194	1.000
13	10979915	1.000
14	10909027	1.000
15	11045788	0.999
16	10796238	0.999
17	10975531	0.999
18	11008441	0.999
19	10935178	0.998
20	10971468	0.998

The second usage of the model could be a structural decision tree of all the predictors (full decision tree in appendix). Hence the decision makers can better understand each significant variable and the relationship and hierarchy of the predictors.

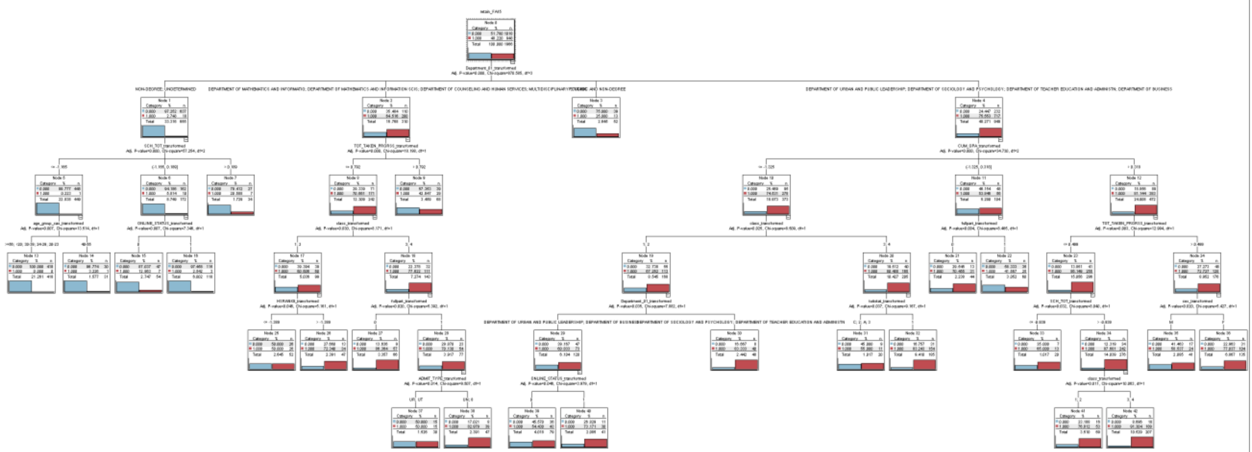


A remarkable variable that draw our attention is the Undetermined under college. As the result of the model indicates, the chance of a student that is not retained will increase significantly if he or she is assigned as undetermined. Future research will be done to find out more information about the correlation between undetermined and retention.

We would like to separate the students that did not return to UNT Dallas as drop out of college or transfer to other institutions in the future project to better understand students' retention.

Appendix

Result from SPSS



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1232.363	76	.000
	Block	1232.363	76	.000
	Model	1232.363	76	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1490.599[a]	.466	.621

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table

			Predicted		
			retain_FA15		Percentage Correct
Observed			0.0	1.0	
Step 1	retain_FA15	0.0	746	272	73.3
		1.0	74	874	92.2
	Overall Percentage				82.4

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SCH_TOT_transformed	.026	.858	.001	1	.976	1.026
	TOT_TAKEN_PRGRSS_transformed	-.599	.139	18.446	1	.000	.549
	UNT_TAKEN_PRGRSS_transformed	-.324	.145	5.007	1	.025	.723
	HSRANK9_transformed	-.125	.149	.702	1	.402	.883
	age_transformed	.544	.276	3.896	1	.048	1.723
	CUM_GPA_transformed	.499	.155	10.377	1	.001	1.646
	TOTALSCHDAL_transformed	1.069	.862	1.537	1	.215	2.913
	sex_transformed(1)	-.188	.153	1.500	1	.221	.829
	class_transformed			27.810	4	.000	
	class_transformed(1)	.234	.927	.064	1	.800	1.264
	class_transformed(2)	-1.281	.424	9.114	1	.003	.278
	class_transformed(3)	-.673	.272	6.131	1	.013	.510
	class_transformed(4)	.254	.197	1.661	1	.197	1.289
	tuitstat_transformed			5.987	5	.308	
	tuitstat_transformed(1)	-1.724	1.375	1.574	1	.210	.178
	tuitstat_transformed(2)	-20.006	11804.185	.000	1	.999	.000
	tuitstat_transformed(3)	-.993	.719	1.907	1	.167	.371
	tuitstat_transformed(4)	.113	.362	.097	1	.756	1.119
	tuitstat_transformed(5)	-.611	.332	3.385	1	.066	.543
	Ethnicity_IPEDS_transformed			11.982	8	.152	